REPORT

A Geometric Analysis of Gaussian Elimination. I*

George Poole and Larry Neal
Mathematics Department
East Tennessee State University
Johnson City, Tennessee 37614-0002

Submitted by David H. Carlson

ABSTRACT

The algorithm known as Gaussian elimination (GE) is fully understood in an exact-arithmetic environment. But in the finite-precision environment of computers, a full understanding of GE has been somewhat elusive. Heretofore, the analysis of this popular and important algorithm has been primarily from a numerical perspective. This paper seeks to analyze GE from a geometric perspective, and by so doing, (1) confirm the classical numerical analysis and (2) demonstrate a new level of understanding through the Euclidean geometry of GE.

1. INTRODUCTION

One of the most popular and useful algorithms of all time designed to solve systems of linear equations is the one known to almost everyone as Gaussian elimination (GE). "If numerical analysts understand anything, surely it must be Gaussian elimination. This is the oldest and truest of algorithms." These words were penned with tongue in cheek by Trefethen [9] in a 1985 article entitled "Three Mysteries of Gaussian Elimination."

This is the first of two papers written from a geometric perspective to unravel some of the interesting mysteries of this remarkable, yet not fully understood algorithm. The algorithm has two phases. The first phase at-

*The contents of this paper were first presented at the Southeast Regional MAA Meeting held in Savannah, Ga. in April 1987. The contents are copyrighted by *Linear Algebra Appl.* and the authors.

LINEAR ALGEBRA AND ITS APPLICATIONS 149:249–272 (1991)

655 Avenue of the Americas, New York, NY 10010 0024-3795/91/$3.50

tempts to transform the original system into an equivalent upper triangular system. This phase is often called the *sweepout phase* (SWOP). The second phase recovers the solution vector by methodically computing the components one at a time, in reverse order. This phase is often referred to as the *back-substitution phase* (BSP).

In the analysis of this algorithm, and in the desire to control errors occurring during the sweepout phase, analysts have extensively studied stability/instability, the growth factor (growth of elements during the SWOP), backward-error analysis, ill-conditioning and condition number, partial pivoting (PP), row scaling followed by partial pivoting (SPP), and complete (or total) pivoting (CP). This list is representative, but by no means complete. It is not our purpose to survey the literature of these topics; we refer the reader to the excellent reference works contained in [2], [3], [4], [7], [8], and [11].

Our goal will be achieved if we can demonstrate through geometric analysis a new level of understanding of GE which is in concert with previous conclusions based on numerical analysis. In this first paper we shall address only the topics of PP, SPP, and CP. The other topics mentioned above will be addressed in the sequel [5], and in [6].

In this paper, we shall use small-dimension examples with specified decimal precision. Neither the size of these linear systems nor their limited precision should diminish the understanding they lend to GE, nor detract from the importance of the conclusions which may be drawn from them. Analogous examples of higher dimension and extended precision can be easily constructed by the reader who fully understands the geometric analysis presented in this paper.

In Section 2 of this paper, we shall use the familiar geometry of hyperplanes to present the geometries of both the SWOP and the BSP of GE. These geometries will then be used to demonstrate some interesting implications regarding pivoting strategies based solely on the magnitude of elements.

In Section 3 we shall present the geometry of classical pivoting strategies and show why they are generally successful in obtaining acceptable computed solutions. We shall also show what causes them to fail from time to time, and how these failures might be anticipated. As a consequence, techniques for generating test matrices will be described in this paper, and in [5].

In Section 4 we shall review some conclusions drawn from the geometric analysis of Gaussian elimination. Remarks and recommendations regarding improved pivoting strategies are deferred until the second paper [5].

All numerical analysts are indebted to James Wilkinson for the level of understanding he brought to GE. This paper is dedicated to his fond memory, and to all those who strive to understand mathematics from a geometric perspective.

2. THE GEOMETRY OF GAUSSIAN ELIMINATION

In this section we shall consider various geometric concepts associated with GE, including both the SWOP and the BSP.

To fully understand the method of GE, it is necessary to gain an appreciation of the role played by each one of the entries in the augmented matrix $[\mathbf{A}\ \mathbf{b}]$ corresponding to the linear system $\mathbf{Ax} = \mathbf{b}$.

Consider the following equation (hyperplane H) in R^n:

$$H: a_1x_1 + a_2x_2 + \cdots + a_nx_n = b. \tag{2.1}$$

Assume that $|a_i| > |a_k|$ for all $k \neq i$. Then the arithmetic of direction cosines implies that H is more nearly orthogonal to the x_i-axis than to any of the other axes. Stated another way, if one coefficient in H is greater in magnitude than all other coefficients, then H intersects the corresponding axis more "crisply" than any of the others. The importance of this fundamental observation relative to GE will soon be demonstrated. As a simple example of this geometry of hyperplanes, consider Figure 1, where the hyperplane $L: a_1x_1 + a_2x_2 = b$ satisfies $|a_1| > |a_2|$. The relative magnitude of a_1 to a_2 (using direction cosines) would indicate that the normal vector (a_1, a_2) is more nearly parallel to the x_1-axis than to the x_2-axis. Hence, L is more nearly orthogonal to the x_1-axis than to the x_2-axis.

Now consider the following system of equations in R^2:

$$\begin{aligned} L_1&: 5.6x_1 - x_2 = 23, \\ L_2&: 14x_1 + 80x_2 = 470. \end{aligned} \tag{2.2}$$

The solution of the system is (5, 5), as shown in Figure 2.

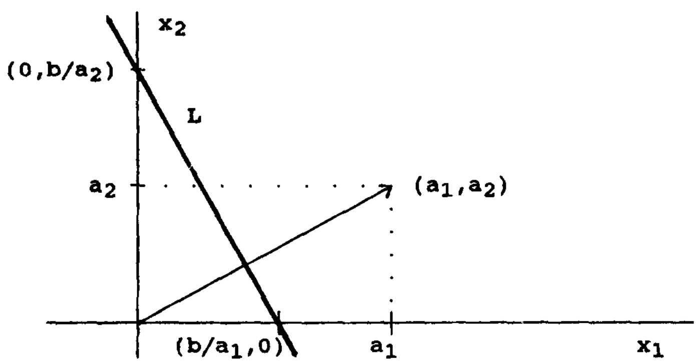

FIG. 1.

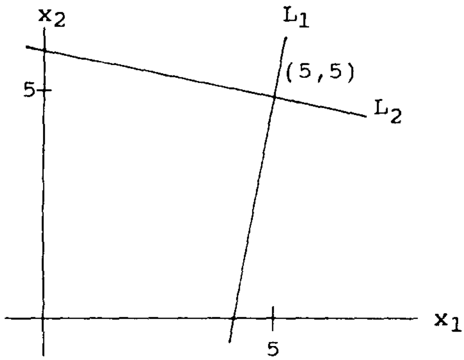

FIG. 2. Graph of the equations (2.2). $L_1: 5.6x_1 - x_2 = 23$; $L_2: 14x_1 + 80x_2 = 470$. L_i is almost perpendicular to axis x_i for $i = 1, 2$. The system is diagonally dominant.

By applying the PP strategy to the system (2.2) using 2-digit arithmetic, phase 1 of GE yields the "equivalent" triangular system (2.3):

$$\begin{aligned} L_2: 14x_1 + 80x_2 &= 470, \\ L_1': \qquad -33x_2 &= -170. \end{aligned} \tag{2.3}$$

Here the equations have been reordered by the PP strategy according to the magnitude of the elements in column 1 of (2.2). The subsequent SWOP leaves the new first equation and its corresponding line L_2 unaltered, while the second equation (L_1) has been replaced by one whose corresponding line L_1' is parallel to the x_1-axis. As indicated by the graph in Figure 3, the potentially catastrophic consequences are realized only after the BSP of GE is completed.

That is, while line L_1 was replaced by one parallel to the x_1-axis, line L_2 remains dangerously close to parallel to the x_1-axis. The consequence is that a small, unavoidable error in the finite-precision computation of the second component of the solution vector will be magnified during the BSP, leading to greater error in the first component. The resulting "solution" may be beyond recognition. In fact, for this ordering of the system (2.3), 2-digit arithmetic yields the computed solution (3.6, 5.2). That is, a relative error of 4% in x_2 is magnified by back-substitution into a relative error of 28% in x_1.

Had one chosen to maintain the original order of the two equations in (2.2) and pivot about the coefficient 5.6 (against the PP strategy, but in

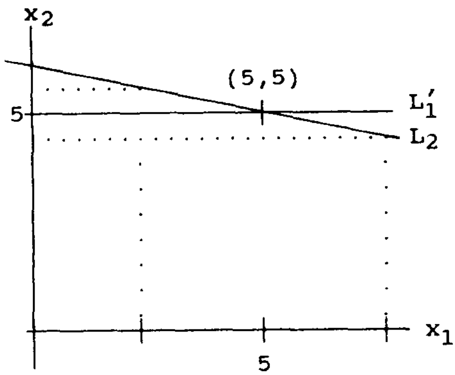

FIG. 3. Graph of the equations (2.3). $L_2: 14x_1 + 80x_2 = 470$; $L_1': -33x_2 = -170$. PP exchanges L_1 and L_2 before the SWOP transforms L_1 into L_1', a new line which is parallel to the x_1-axis. The BSP magnifies any error in x_2.

agreement with explicit row scaling followed by PP), then the system (2.2) would have been replaced by the system

$$\begin{aligned} L_1 &: 5.6x_1 - x_2 = 23, \\ L_2' &: \qquad 83x_2 = 410, \end{aligned} \tag{2.4}$$

whose graph appears in Figure 4. Because L_1 is more nearly perpendicular to the x_1-axis, on replacing L_2 (which is already close to parallel to the x_1-axis) with a line parallel to the x_1-axis, the BSP of GE will not magnify the error in x_2. In fact, the error is reduced. The computed solution in this second ordering is (5.0, 4.9).

Now consider the following triangular system (2.5) in R^3, and the related triangular system (2.6). The solution to each system is (2, 2, 2). We shall use these two systems to demonstrate in R^3 the geometric consequences of hyperplane orientation in the analysis of error during the BSP of GE:

$$\begin{aligned} H_1 &: x + 3y + 2z = 12, \\ H_2 &: \qquad y + 2z = 6, \\ H_3 &: \qquad \tfrac{1}{3}z = \tfrac{2}{3} \end{aligned} \tag{2.5}$$

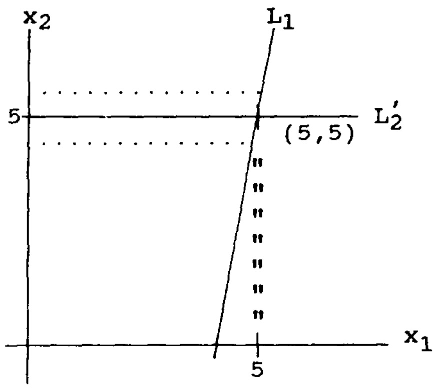

FIG. 4. Graph of the equations (2.4). $L_1: 5.6x_1 - x_2 = 23$; $L_2': 83x_2 = 410$. Implicit scaling with PP recognizes diagonal dominance in the system (2.2). The SWOP replaces L_2 by L_2'. The BSP does not magnify error in x_2 when x_1 is computed.

and

$$\begin{aligned} H_*&: 3x + 2y + z = 12, \\ H_2&: \qquad y + 2z = 6, \\ H_3&: \qquad\quad \tfrac{1}{3}z = \tfrac{2}{3}. \end{aligned} \tag{2.6}$$

We shall use the system (2.5) to illustrate how poorly oriented hyperplanes in the triangular system resulting from the SWOP of GE contribute to the unnecessary growth of error during back-substitution. We shall then use the system (2.6) to illustrate how well-oriented hyperplanes in the triangular system can actually reduce any error which may have accrued earlier due to finite-precision calculations in the SWOP of GE.

The hyperplane H_i in the triangular system resulting from the SWOP of GE will be called *poorly oriented with respect to the* x_i*-axis* if at least one off-diagonal entry a_{ij} satisfies $|a_{ij}| > |a_{ii}|$. Otherwise, H_i will be called *well oriented with respect to the* x_i*-axis*.

In either system (2.5) or (2.6), the solution is obtained as follows: Geometrically, $(2,2,2)$ is the point of intersection of the three hyperplanes given in (2.5) or (2.6). The BSP begins by solving the equation of H_3 for z. Since its x and y coefficients are both zero, H_3 is a plane orthogonal to the z-axis. Therefore, the geometric interpretation in R^3 of the algebraic process of solving the equation $\frac{1}{3}z = \frac{2}{3}$ for z is to locate the point of intersection of the z-axis with the plane H_3 parallel to the xy-plane, namely the point

$(0,0,2)$. The BSP continues by solving the equation of H_2 for y with z set equal to 2. This yields $y=2$. The geometric interpretation of the second step of the BSP is this: from the point where the z-axis intersects H_3, namely $(0,0,2)$, follow the line in H_3 which is parallel to the y-axis until it intersects the plane described by H_2. This point of intersection is $(0,2,2)$ in the yz-plane. Since H_2 and H_3 are both parallel to the x-axis, these two planes intersect in a line which is parallel to the x-axis and orthogonal to the yz-plane. The final step in the BSP is to solve the equation of H_1 (or H_*) for x using values $y=2$ and $z=2$ obtained from the first two steps. This final step yields $x=2$. The geometric interpretation of this step is as follows: from the point $(0,2,2)$, follow the line parallel to the x-axis until it intersects the plane H_1 (or H_*). This line intersects H_1 (or H_*) at the solution point $(2,2,2)$.

We now examine the geometric consequences of a slight error in the first step of the back-substitution phase. Suppose that the numeric solution of $\frac{1}{3}z=\frac{2}{3}$ is $z=2+\epsilon_z$ for some small error ϵ_z. Figure 5 depicts the situation in which $\epsilon_z>0$. Then the BSP begins at the intersection of the z-axis with the horizontal plane H_3' described by the equation $z=2+\epsilon_z$. Thus, from the point $(0,0,2+\epsilon_z)$, the BSP follows the line in H_3' which is parallel to the y-axis until that line intersects the plane H_2. To find this point of intersec-

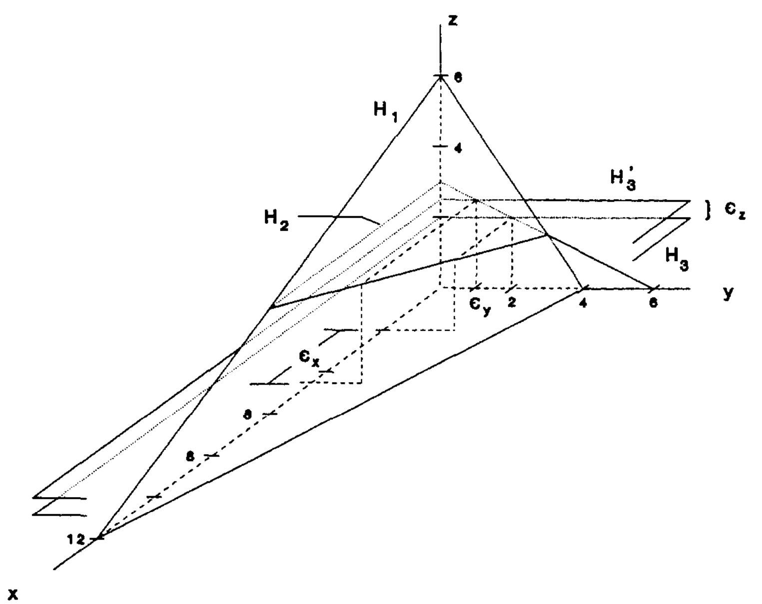

FIG. 5.

tion, solve the equation of H_2 for y using $z = 2 + \epsilon_z$. This yields $y = 2 - 2\epsilon_z$. The poor orientation of H_2 with respect to the y-axis has magnified the initial error ϵ_z so that the error in y is $\epsilon_y = -2\epsilon_z$. The intersection of H_2 and H_3' is the line parallel to the x-axis through the point $(0, 2 - 2\epsilon_z, 2 + \epsilon_z)$. Using back-substitution to find the point of intersection of this line with the plane H_1, we must solve $x + 3(2 - 2\epsilon_z) + 2(2 + \epsilon_z) = 12$ for x. This yields $x = 2 + 4\epsilon_z$, which implies that the point of intersection of the three planes H_1, H_2, and H_3' is $(2 + 4\epsilon_z, 2 - 2\epsilon_z, 2 + \epsilon_z)$. Note that since H_1 is more nearly orthogonal to the y-axis, the final step in the BSP has magnified the initial error so that the error in the x-component of the solution of the system is $\epsilon_x = 4\epsilon_z$. The compound effect of two poorly oriented hyperplanes in a triangular system of order 3 is potentially a serious matter. However, the effect of $n - 1$ poorly oriented hyperplanes in a triangular system of order n is potentially far more serious.

In contrast, the system (2.6) is a triangular system with the same second and third equations as the system (2.5). H_2 is still poorly oriented. However, H_1 has been replaced with H_*, a plane more nearly orthogonal to the x-axis

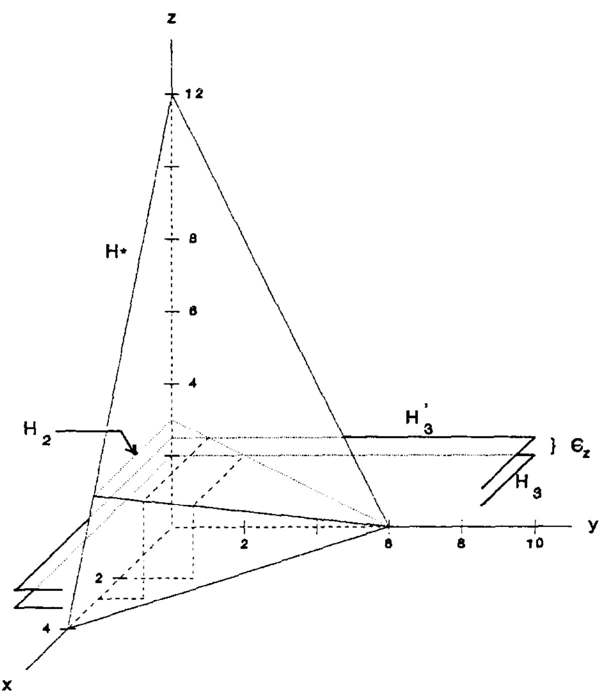

FIG. 6.

than to either of the other axes. As before, assume in this second system that a slight error is made at the initial step of the BSP, say $z = 2 + \epsilon_z$. Figure 6 depicts the situation in which $\epsilon_z > 0$. As in the previous example, the second step of the BSP starts at $(0, 0, 2 + \epsilon_z)$ and locates the next point $(0, 2 - 2\epsilon_z, 2 + \epsilon_z)$. But since H_* is more nearly orthogonal to the x-axis than was H_1 in the system (2.5), the projection from $(0, 2 - 2\epsilon_z, 2 + \epsilon_z)$ intersects H_* at a point whose x-coordinate is much closer to 2 than is the x-coordinate of that same projection onto the plane H_1. To verify this algebraically, solve the equation H_* for x with $z = 2 + \epsilon_z$ and $y = 2 - 2\epsilon_z$. This yields $x = 2 + \epsilon_z$. Thus, the magnification of the initial error ϵ_z due to the poor orientation of H_2 has been overcome by the subsequent reduction of error in x due to the favorable orientation of H_*.

To underscore the point that the orientation of hyperplanes in the triangular system generated by the SWOP can be critical, consider the system (2.7) in which $[10, 10, 10, 10, 10, 10]^t$ is the solution and satisfies the first five equations *exactly*. Each hyperplane associated with the first five equations is nearly orthogonal to the last axis of R^6, an unfortunate situation:

$$\begin{bmatrix} 1 & 10 & 20 & 30 & 40 & 50 & 1510 \\ & 1 & 10 & 20 & 30 & 40 & 1010 \\ & & 1 & 10 & 20 & 30 & 610 \\ & & & 1 & 10 & 20 & 310 \\ & & & & 1 & 10 & 110 \\ & & & & & * & * \end{bmatrix}. \tag{2.7}$$

If $x_6 = 10.001$ instead of 10, the computed solution of (2.7) is

$$x_6 = 10.001, \qquad x_5 = 9.99, \qquad x_4 = 10.08,$$

$$x_3 = 9.37, \qquad x_2 = 14.96, \qquad x_1 = -29.05.$$

By assuming a *very small relative error* of 0.01% in the computation of x_6 and computing the remaining five components in *exact arithmetic* (so that no further computational error is introduced), the implications of *poor* hyperplane orientation in the BSP are clear. There can be no misunderstanding that the errors in the last five computed components of the solution are geometric (not numeric). Poor orientation of hyperplanes in the triangular system can be a serious matter. That is, the BSP of GE can be highly unstable, even though the first phase may have been stable due to a pivoting strategy designed to control the growth of error during the SWOP.

Now consider the following triangular system (2.8) in which $[10, 10, 10, 10, 10, 10]^t$ is the solution and satisfies the first five equations *exactly*. In contrast to the triangular system (2.7) above, the first five hyperplanes are more nearly orthogonal to the first five axes of R^6, respectively:

$$\begin{bmatrix} 100 & 70 & 50 & 30 & 10 & 1 & 2610 \\ & 100 & 60 & 40 & 20 & 1 & 2210 \\ & & 100 & 50 & 30 & 1 & 1810 \\ & & & 100 & 40 & 1 & 1410 \\ & & & & 100 & 1 & 1010 \\ & & & & & * & * \end{bmatrix}. \tag{2.8}$$

If $x_6 = 20$ instead of 10, the computed solution of the system (2.8) is

$$x_6 = 20.0, \qquad x_5 = 9.9, \qquad x_4 = 9.94,$$

$$x_3 = \ 9.96, \qquad x_2 = 9.968, \qquad x_1 = 9.9704.$$

By assuming a *large relative error* of 100% in the computation of x_6 and computing the remaining five components in *exact arithmetic* (so that no further computational error is introduced), the implications of *good* hyperplane orientation in the BSP are clear. In this particular example there is, in fact, a componentwise geometric convergence occurring during the BSP. That is, after the first component is determined, each subsequent computed component is more nearly accurate than previously computed components.

REMARK. This is the proper time to make the point that a good pivoting strategy should address possible numerical instability in both phases of GE. In order to do so, the pivoting strategy should reflect a careful analysis of the orientation of the hyperplanes (corresponding to the equations of a linear system) with respect to the coordinate axes before either phase of GE is applied. In fact, after the SWOP is completed, no further opportunities exist to control errors caused by poor orientation of hyperplanes during the BSP.

We now return to the geometry of the SWOP. When a pivot element is selected for position (k, k) of a linear system of order n and the subsequent sweepout step is completed, it will be convenient in the discussions which follow to say that the hyperplanes corresponding to rows $k+1$ through n have been *rotated* about the solution point into hyperplanes parallel to the x_k-axis. For example, when the system (2.2) is transformed into the equivalent system (2.4), we shall say that the line L_2 is rotated about (5, 5) into the line L_2', which is parallel to the x_1-axis. The operation of adding a scalar multiple of one equation to a second equation to eliminate the x_1-coefficient of the second equation is not an operation that results in a true rotation. That is, although the second line L'_2 in (2.4) is parallel to the x_1-axis, it is not a result of rotating the line L_2 toward the x_1-axis in the classical sense. But that is a vivid phrase, and we shall use it without further caution.

Before considering some examples in the next section to illustrate implications of the geometric analysis regarding the classical pivoting strategies, we shall review the geometric concepts associated with GE.

Consider the following arbitrary system of linear equations (2.9) in R^3, which we assume to have a unique solution:

$$\begin{aligned} H_1 &: a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1, \\ H_2 &: a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = b_2, \\ H_3 &: a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = b_3. \end{aligned} \tag{2.9}$$

Without regard to any kind of pivoting strategy, the two steps of the SWOP transform (2.9) into the following equivalent systems, respectively:

$$\begin{aligned} H_1 &: a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1, \\ H'_2 &: \qquad\qquad a'_{22}x_2 + a'_{23}x_3 = b'_2, \\ H'_3 &: \qquad\qquad a'_{32}x_2 + a'_{33}x_3 = b'_3. \end{aligned} \tag{2.10}$$

and

$$\begin{aligned} H_1 &: a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1, \\ H'_2 &: \qquad\qquad a'_{22}x_2 + a'_{23}x_3 = b'_2, \\ H''_3 &: \qquad\qquad\qquad\qquad a''_{33}x_3 = b''_3. \end{aligned} \tag{2.11}$$

Suppose $\mathbf{X}$ is the solution to (2.9). That is, $\mathbf{X}$ is the point of intersection of the three hyperplanes H_1, H_2, and H_3. After the first step of the SWOP, (2.9) has been transformed to the equivalent system (2.10). Here the first plane H_1 is unchanged. But H_2 and H_3 have been rotated about the solution $\mathbf{X}$ into planes H'_2 and H'_3, respectively, two planes which are parallel to the x_1-axis. That is, if L denotes the line of intersection of planes H'_2 and H'_3, then L is parallel to the x_1-axis and L intersects H_1 at the solution vector $\mathbf{X}$.

Upon completion of the second and final step of the SWOP, the system (2.10) is transformed into the equivalent system (2.11). Here the first and second planes H_1 and H_2' are unchanged. But H_3' is rotated about the line L into the plane H_3'', which is parallel to both the x_1-axis and the x_2-axis (and orthogonal to the x_3-axis).

Now consider the three planes H_1, H_2', and H_3''. These are the planes used during the BSP of GE to determine the components of the solution vector **X**. In order to avoid instability of the type illustrated by Figures 3 and 5 during the BSP, one would prefer that the system (2.11) and its corresponding planes H_1, H_2', and H_3'' satisfy the following relationships. Since L denotes the line of intersection of H_2' and H_3'', then ideally, H_1 should be as close to orthogonal to the line L as possible in order for the intersection of H_1 and L to be as crisp as possible. Since L is parallel to the x_1-axis, this condition is equivalent to requiring that H_1 be as nearly orthogonal to the x_1-axis as possible.

Also, H_2' should not be nearly parallel to the plane H_3'', in order to avoid the problems suggested by Figures 3 and 5. Since H_2' is parallel to the x_1-axis, H_2' should be as nearly orthogonal to the x_2-axis as possible in order to insure as crisp an intersection of these planes as possible.

In summary, the geometry of GE is this: the sweepout phase replaces selected hyperplanes in the original linear system by new hyperplanes which are parallel to selected axes, forming a triangular system. Theoretically, this new set of hyperplanes intersects at the same point as the original set of hyperplanes. During the SWOP, the rearrangement of equations (and possibly unknowns) should be conducted so that prior to the BSP, the ith hyperplane is as nearly orthogonal to the x_i-axis as possible. If after the SWOP the ith hyperplane is actually orthogonal to the ith coordinate axis for $i = 1, 2, \dots, n$, then phase 1 of GE produced a diagonal system, the ideal situation for the BSP. This situation is ideal not only in that few computations are required to determine the solution, but more importantly, in that the intersection of all hyperplanes is as crisp as possible. The mutual orthogonality of these hyperplanes has the effect of isolating the calculation of x_i from the calculation of x_j for $i \neq j$, thus completely avoiding instability during the BSP.

Although the situation just described is ideal for the BSP of GE, unfortunately one must monitor and control the errors being generated during the SWOP. But likewise, one cannot be consumed by the task of reducing errors during the SWOP and ignore the potential problems in the BSP. It is worth saying twice (here and in Section 4) that the *total error* in the computed solution of a linear system by GE can be minimized only by controlling the errors stemming from both the SWOP and the BSP. We shall illustrate this point further in Section 3.

From these geometric considerations, one can easily see why a strictly diagonally dominant system (determined by explicit row scaling and PP) presents few problems during the implementation of GE. The algorithm is "doubly stable" in this case. That is, each step of the SWOP is stable due to the relatively large divisors on the diagonal, and the BSP is stable due to the excellent orientation of the hyperplanes in the triangular system leading to crisp intersections. The system given in (2.2) illustrates this point, as does Example 3.2 in the following section.

3. CLASSICAL PIVOTING STRATEGIES AND THE BSP OF GE

In this section we shall illustrate by geometric example why classical pivoting strategies, including scaling, work as well as they do. But we shall also illustrate how and why these strategies sometimes fail to produce acceptable computed solutions.

It has been demonstrated many times over in the literature that in the selection of a pivot, one should choose a relatively large element to minimize errors in the SWOP generated by divisions in which the quotient exceeds one in magnitude. However, most linear systems could actually tolerate quotients whose magnitudes exceed one significantly, although the degree depends on several factors. Some of these are discussed in [5] and [6]. Indeed, when implicit scaling [3] is implemented, the selected pivots quite often produce scales with magnitudes in excess of unity. Nevertheless, much error analysis of GE assumes that the quotients involved in the SWOP have magnitudes less than or equal to one [3, 4, 8, 11].

Although we shall discuss the subjects of ill-conditioning and condition number in the sequel [5], it is necessary to mention them in the context of the LU decomposition to better explain the point made above regarding the magnitudes of quotients in the SWOP of GE.

For some permutation matrices $\mathbf{P}$ and $\mathbf{Q}$ (determined by the method of selecting pivots), an LU decomposition of $\mathbf{PAQ}$ takes the form

$$\begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ s_{21} & 1 & 0 & 0 & \cdots & 0 \\ s_{31} & s_{32} & 1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ s_{n1} & s_{n2} & s_{n3} & s_{n4} & \cdots & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ 0 & u_{22} & u_{23} & \cdots & u_{2n} \\ 0 & 0 & u_{33} & \cdots & u_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & u_{nn} \end{bmatrix}. \tag{3.1}$$

Prior to interchanging rows and/or columns based on some pivoting strategy, some analysts have advocated "scaling" the system $\mathbf{Ax} = \mathbf{b}$ on the "front end" before applying GE. That is, GE is applied to the system $(\mathbf{D}_1\mathbf{A}\mathbf{D}_2)\mathbf{y} = \mathbf{D}_1\mathbf{b}$ [where $\mathbf{D}_1$ and $\mathbf{D}_2$ are diagonal matrices and $\mathbf{y} = (\mathbf{D}_2)^{-1}\mathbf{x}$]. Regarding the subject of scaling, it is important to note that in the absence of row and/or column interchanges, scaling the rows or columns of $\mathbf{A}$ by powers of the machine base in no way affects the computed solution [1]. The only way different computed solutions are obtained by GE is to alter the order of the arithmetic computations by rearranging rows and/or columns of the coefficient matrix. This rearrangement is "driven" by the choice of scaling on the front end and the method of pivot selection. In this section we shall explain what row scaling may or may not accomplish in "driving" the algorithm toward an acceptable computed solution. However, we shall discuss the subject of both row and column scaling more thoroughly in [6].

Given the linear system $\mathbf{Ax} = \mathbf{b}$ of order n, there are at most $(n!)^2$ possible *LU* decompositions. Consequently, using GE there are at most $(n!)^2$ different computed solutions to the system $\mathbf{Ax} = \mathbf{b}$. In a finite-precision environment, these computed solutions could all be different. Among the $(n!)^2$ possible computed solutions, rarely are the solutions obtained from the classical pivoting strategies the best possible [5]. However, empirical evidence suggests that solutions computed using classical pivoting strategies are acceptable far more often than not.

Preliminary testing reveals that often the smallest absolute or relative error occurs in those computed solutions in which the "scales" (elements below the diagonal of $\mathbf{L}$) exceed one in magnitude. That is, the condition number of $\mathbf{L}$ can be larger than that realized when PP or CP is used as the pivoting strategy. We shall demonstrate geometrically why this can be true.

Recall that the steps of GE [3], in the language of the *LU* decomposition, are (1) compute $\mathbf{L}$ and $\mathbf{U}$, (2) forward-solve the lower triangular system $\mathbf{Ly} = \mathbf{b}$, and (3) backward-solve the upper triangular system $\mathbf{Ux} = \mathbf{y}$ (we may omit permutations and scalings).

Observations about Partial Pivoting. PP attempts to control errors in $\mathbf{U}$ by avoiding divisions of large elements by small elements during the swop. Before we present some examples, consider the following observations regarding PP and (3.1):

(1) The topmost rows of $\mathbf{L}$ and $\mathbf{U}$ usually contain less computational error than the bottommost rows.

(2) $\mathbf{L}$ is *diagonally maximal* (($|s_{ij}| \leqslant |s_{ii}|$ *for all* $i \neq j$), implying good hyperplane orientation in $\mathbf{Ly} = \mathbf{b}$.

(3) Although PP provides some positive control over error in the elements of $\mathbf{U}$, there is *no explicit control* over the hyperplane orientation in $\mathbf{Ux} = \mathbf{y}$ to avoid the kind of instability revealed in Figures 3 and 5.

(4) Compared to Cond(**A**), Cond(**L**) is relatively small while Cond(**U**) may be quite large.

(5) Determining **y** in **Ly** = **b** is stable.

(6) Determining **x** in **Ux** = **y** may be unstable.

In the examples to follow, all numbers are chosen so that no immediate representation error is injected into the system when the system is machine stored.

EXAMPLE 3.1. In the example below (presented as an augmented system), PP with 3-digit arithmetic is used to compute **L**, **U**, and the solution. The condition numbers (obtained from MATLAB [12]) are presented for observations. Note that PP generates a well-conditioned **L**, while the matrix **U** contains poorly oriented hyperplanes for the BSP:

$$[\mathbf{A} \quad \mathbf{b}] = \begin{bmatrix} 19 & 45 & 13.5 & -10 & 675 \\ 4 & 28 & 57 & 22 & 1110 \\ -6 & 3.5 & 7 & -21 & -165 \\ 10 & 7 & -2.5 & 0 & 145 \end{bmatrix}.$$

The exact solution is $[10, 10, 10, 10]^t$, and we have

$$\overset{\mathbf{L}}{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.212 & 1 & 0 & 0 \\ -0.316 & 0.957 & 1 & 0 \\ 0.526 & -0.903 & -0.970 & 1 \end{bmatrix}} \overset{\mathbf{U}}{\begin{bmatrix} 19 & 45 & 13.5 & -10 \\ 0 & 18.5 & 54.1 & 24.1 \\ 0 & 0 & -40.5 & -47.5 \\ 0 & 0 & 0 & -19.0 \end{bmatrix}},$$

$$\mathrm{Cond}(\mathbf{A}) = 19.6, \qquad \mathrm{Cond}(\mathbf{L}) = 4.09, \qquad \mathrm{Cond}(\mathbf{U}) = 28.1.$$

The computed solution is

$$x_4 = 9.89,$$

$$x_3 = 10.0,$$

$$x_2 = 10.2,$$

$$x_1 = 9.47.$$

EXAMPLE 3.2. For the system presented in Example 3.1, notice that the four equations may be rearranged into a strictly diagonally dominant system.

Explicitly scaling each row by the element of largest magnitude in that row is an effective strategy for identifying strictly diagonally dominant systems. Consequently, if row scaling were employed before the SWOP was begun, PP would identify as the ith pivot row that hyperplane most nearly orthogonal to the x_i-axis (with implicit row scaling as advocated in [3]; the scales are stored, but scaling entire rows is avoided). Below, $\mathbf{PA} = \mathbf{LU}$, where $\mathbf{Pb} = \mathbf{b}' = [145, 675, 1110, -165]^t$:

$$\overset{\mathbf{L}}{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1.9 & 1 & 0 & 0 \\ 0.4 & 0.795 & 1 & 0 \\ -0.6 & 0.243 & 0.0241 & 1 \end{bmatrix}} \overset{\mathbf{U}}{\begin{bmatrix} 10 & 7 & -2.5 & 0 \\ 0 & 31.7 & 18.3 & -10 \\ 0 & 0 & 43.5 & 30.0 \\ 0 & 0 & 0 & -19.3 \end{bmatrix}},$$

$$\operatorname{Cond}(\mathbf{A}) = 19.6, \qquad \operatorname{Cond}(\mathbf{L}) = 7.34, \qquad \operatorname{Cond}(\mathbf{U}) = 6.38.$$

The computed solution is

$$x_4 = 10.0,$$

$$x_3 = 10.0,$$

$$x_2 = 9.97,$$

$$x_1 = 10.0.$$

Now consider a situation where scaling hinders the implementation of PP.

EXAMPLE 3.3. In the example below (presented as an augmented system), the first step of the SWOP has been completed. PP with implicit row scaling and 3-digit arithmetic is used to compute $\mathbf{L}$ and $\mathbf{U}$, and the solution $\mathbf{x}$. The computed solution in this case is not as good as the one computed *without* row scaling as presented in Example 3.4. The reason is that neither row scaling nor partial pivoting is effective in ferreting out good hyperplane orientations when the large elements of the coefficient matrix cluster in one or a few columns. See Remarks 4.3 in the next section.

Let

$$[\mathbf{A} \quad \mathbf{b}] = \begin{bmatrix} 10 & 5 & 2 & 0 & 170 \\ 0 & 1 & 16 & 17 & 340 \\ 0 & 3 & 0 & 52 & 550 \\ 0 & 1 & 54 & 65 & 1200 \end{bmatrix}.$$

The exact solution is $[10, 10, 10, 10]^t$, and we have

$$\overset{\mathbf{L}}{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 3 & 1 & 0 \\ 0 & 1 & -0.792 & 1 \end{bmatrix}} \overset{\mathbf{U}}{\begin{bmatrix} 10 & 5 & 2 & 0 \\ 0 & 1 & 16 & 17 \\ 0 & 0 & -48 & 1 \\ 0 & 0 & 0 & 47.2 \end{bmatrix}},$$

$$\text{Cond}(\mathbf{A}) = 137, \qquad \text{Cond}(\mathbf{L}) = 16.5, \qquad \text{Cond}(\mathbf{U}) = 65.7.$$

The computed solution is

$$x_4 = 10.3,$$
$$x_3 = 10.0,$$
$$x_2 = 5.0,$$
$$x_1 = 12.5.$$

EXAMPLE 3.4. The computed solution to the system given in Example 3.3 is now obtained using 3-digit arithmetic and PP, but with no row scaling. Although not good, the computed solution without row scaling is distinctly better. Below, $\mathbf{PA} = \mathbf{LU}$, where $\mathbf{Pb} = \mathbf{b}' = [170, 550, 1200, 340]^t$:

$$\overset{\mathbf{L}}{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0.333 & 1 & 0 \\ 0 & 0.333 & 0.296 & 1 \end{bmatrix}} \overset{\mathbf{U}}{\begin{bmatrix} 10 & 5 & 2 & 0 \\ 0 & 3 & 0 & 52 \\ 0 & 0 & 54 & 47.7 \\ 0 & 0 & 0 & -14.4 \end{bmatrix}},$$

$$\text{Cond}(\mathbf{A}) = 137, \qquad \text{Cond}(\mathbf{L}) = 1.59, \qquad \text{Cond}(\mathbf{U}) = 116.$$

The computed solution is

$$x_4 = 10.1,$$
$$x_3 = 9.96,$$
$$x_2 = 8.33,$$
$$x_1 = 10.8.$$

Before presenting our last examples, which compare PP with CP, we present the following observations about CP.

Observations about Complete Pivoting. CP attempts to control errors in **U** by avoiding divisions of large elements by small elements during the swop. However, CP also leads to a good orientation of hyperplanes in both the **Ly** = **b** and **Ux** = **y** systems.

Relative to the *LU* decomposition given in (3.1) note that:

(1) The topmost rows of **L** and **U** usually contain less computational error than the bottommost rows.

(2) **L** is *diagonally maximal*, implying good hyperplane orientation in **Ly** = **b**.

(3) **U** is also diagonally maximal. In addition to providing some positive control over error in the elements of **U**, CP also leads to good hyperplane orientation in the system **Ux** = **y** to avoid the kind of instability revealed in Figures 3 and 5.

(4) Compared to Cond(**A**), Cond(**L**) is small. Although Cond(**U**) can be quite large, it is generally smaller than it would be if PP were employed.

(5) Determining **y** in **Ly** = **b** is stable.

(6) Determining **x** in **Ux** = **y** is generally stable.

However, one should not assume that good or best possible computed solutions are obtained with CP [5]. Consider the next examples.

Example 3.5. In the system below, PP is applied with 3-digit arithmetic. Note that the computed solution is the exact solution. However, as observed in Example 3.6, when CP is applied the computed solution is not as good. In this case, that is because PP selects a pivot equation (the first equation) which is more nearly orthogonal to its pivot axis than CP selects. CP selects the second equation for the first pivot equation and the third axis for its associated pivot axis. The second equation is less nearly orthogonal to the x_3-axis. $\mathbf{PA} = \mathbf{LU}$, where $\mathbf{Pb} = \mathbf{b}' = [120, 380, 460]^t$:

$$[\mathbf{A} \quad \mathbf{b}] = \begin{bmatrix} 11 & 0 & 1 & 120 \\ 10 & 17 & 19 & 460 \\ 3 & 18 & 17 & 380 \end{bmatrix}.$$

The exact solution is $[10, 10, 10]^t$, and we have

$$\overset{\mathbf{L}}{\begin{bmatrix} 1 & 0 & 0 \\ 0.273 & 1 & 0 \\ 0.909 & 0.944 & 1 \end{bmatrix}} \overset{\mathbf{U}}{\begin{bmatrix} 11 & 0 & 1 \\ 0 & 18 & 16.7 \\ 0 & 0 & 2.3 \end{bmatrix}},$$

$$\text{Cond}(\mathbf{A}) = 34.4, \qquad \text{Cond}(\mathbf{L}) = 3.21, \qquad \text{Cond}(\mathbf{U}) = 14.7.$$

The computed solution is

$$x_3 = 10.0,$$

$$x_2 = 10.0,$$

$$x_1 = 10.0.$$

EXAMPLE 3.6. In this example, CP with 3-digit arithmetic is applied to the linear system presented in Example 3.5. Although the computed solution is reasonable (maybe even acceptable), it is not as good as the one produced from PP, for the reasons given in Example 3.5. We have $\mathbf{PA} = \mathbf{LU}$, where $\mathbf{Pb} = \mathbf{b}' = [460, 120, 380]^t$:

$$\overset{\mathbf{L}}{\begin{bmatrix} 1 & 0 & 0 \\ 0.0526 & 1 & 0 \\ 0.895 & -0.567 & 1 \end{bmatrix}} \overset{\mathbf{U}}{\begin{bmatrix} 19 & 10 & 17 \\ 0 & 10.5 & -0.894 \\ 0 & 0 & 2.29 \end{bmatrix}},$$

$$\operatorname{Cond}(\mathbf{A}) = 34.4, \qquad \operatorname{Cond}(\mathbf{L}) = 2.79, \qquad \operatorname{Cond}(\mathbf{U}) = 16.7.$$

Noting the column interchanges required for CP, the computed solution in the order (left to right) of the computed components is

$$x_2 = 9.61, \qquad x_1 = 9.9, \qquad x_3 = 10.3.$$

Lack of space precludes the presentation of many other interesting and enlightening examples. In particular, we have omitted an example to demonstrate the importance of controlling error in the SWOP. But with the tools provided in this paper, one should be able to construct a simple (or complex) example for which the relentless desire to provide well-oriented hyperplanes in the system $\mathbf{Ux} = \mathbf{y}$, to the exclusion of monitoring the size of scales in $\mathbf{L}$, will prove disastrous. Even more, a full understanding of the geometry behind each phase of GE provides one with the skill to generate example or test matrices for pedagogical purposes and algorithm design.

4. SUMMARY AND CONCLUSIONS

Our goal in this paper has been to share a geometric analysis of GE which, for us, has led to a better understanding of this intriguing algorithm.

From this analysis we believe that the difficulties of implementing GE in a finite-precision environment can be better anticipated and controlled. Furthermore, we believe that better test matrices can be constructed to analyze strategies used in the implementation of GE.

Based on the (Euclidean) geometric analysis of GE, as well as comments and questions from both informal and formal reviewers, we offer the following remarks:

REMARK 4.1. Geometric analysis of GE does not support the use of PP without some additional knowledge about the coefficient matrix or some preconditioning scheme. The BSP can be seriously unstable and lead to an unacceptable computed solution.

REMARK 4.2. Empirical evidence reveals that PP leads to acceptable computed solutions among "most" linear systems. We believe this may be attributed to one or more of the following:

(1) Naturally occurring linear systems are "benign" in that the associated hyperplanes are not nearly parallel to any axis. Trefethen expressed this same view in [10], but used the terminology of "average case" stability based on the "normal" distribution of the elements in the coefficient matrix.

(2) The larger elements of the coefficient matrix are uniformly distributed across the rows and columns, so that PP, while promoting stability during the SWOP, favorably orients the hyperplanes for the BSP (e.g. in diagonally maximal systems).

(3) The order of the computer precision is sufficiently large to postpone ("cover up") the manifestation of any instability which might occur in the BSP. However, as illustrated in (2.7) above, there are situations in which extended precision (even exact arithmetic) employed during the BSP may not protect one from poor decisions made during the SWOP.

(4) Even if PP (or any other strategy) results in a triangular system of hyperplanes which are poorly oriented with respect to coordinate axes, cancellation of error may occur during the BSP. A condition which assures such cancellation is presented in [5].

We believe the recent work of Trefethen [9, 10] supports Remark 4.2 from a statistical perspective. However, with reasonable caution (and minimal expense) there is really no need to gamble that PP will produce an acceptable solution [5].

REMARK 4.3. SPP (explicit row scaling followed by partial pivoting) is generally successful when the larger elements of the coefficient matrix are uniformly distributed across its rows and columns. In this case, SPP recognizes when the system can be rearranged into a diagonally maximal (or diagonally dominant) system. The hyperplane which is most nearly orthogonal to the pivot axis is the one whose normal is most nearly parallel to that axis—that is, the one whose direction cosine with respect to the pivot axis is closest to one in absolute value. The denominator used in calculating the direction cosine associated with each hyperplane is the L_2-norm of the normal to that hyperplane (the L_2-norm of a row vector). Explicit row scaling divides each entry in a row by the L_∞-norm of that row. PP applied after row scaling approximates the process of comparing direction cosines of hyperplanes with respect to the pivot axis. If each row has a dominant entry (as in the case of a diagonally dominant system), the L_∞-norm of each row is a good approximation to the L_2-norm of that row. In this case, PP after row scaling (based on the L_∞-norms of the rows) will select as the pivot hyperplane the one which is most nearly orthogonal to the pivot axis. Thus, for strictly diagonally dominant systems, SPP makes the same pivot decisions that a strategy based on direction cosines would make, because the L_∞-norm in each row of such a system is a good approximation to the corresponding L_2-norm.

On the other hand, when the elements with largest magnitudes in each row happen to cluster in one or a few columns other than the pivot column, SPP does not always select as the pivot hyperplane that one which is most nearly orthogonal to the pivot axis. That is, dividing the ith entry in column k by the L_∞-norm of row i does not always produce a good estimate of the direction cosine of the ith hyperplane with respect to the kth axis [6] unless the system can be rearranged into a strictly diagonally dominant system. Whenever possible, SPP produces such a rearrangement.

Remark 4.4. CP almost always provides "insurance" against instability in the BSP of GE, and by design encourages stability during the SWOP. The geometric analysis indicates that the more benign the linear system, the less effective this strategy becomes (compared to PP). That is, among benign systems, PP produces computed solutions which are as good as those produced by CP. Like PP, the CP strategy is based on the magnitude of elements in the coefficient matrix, and this strategy alone seldom makes the wisest of all possible choices for pivot elements.

Remark 4.5. While monitoring stability during the SWOP, one should note that after this phase of GE is completed, no further opportunities exist to control errors caused by instability during the BSP. Therefore, during each step of the SWOP, it behooves one to consider the future consequences of a pivot decision, as well as the immediate consequences. More succinctly, reduction of total error in the computed solution of a linear system by GE requires one to employ a pivoting strategy which explicitly addresses errors in both the SWOP and the BSP. An efficient strategy designed to do so will be presented in [5].

REMARK 4.6. One may question how much can be learned about GE from studying examples of small systems of linear equations. Nevertheless, embedded in the GE algorithm to solve any large linear system is the unavoidable task of solving a linear system of order 2 which, in turn, is embedded in a linear system of order 3, etc. Data from large examples will be presented in the sequel [5].

REMARK 4.7. The geometry presented here may lead some readers to question how it compares to geometrical pictures presented in texts on this subject. For example, the figure on p. 206 of *Numerical Methods for Science and Engineering* by R. Stanton is a diagram which attempts to explain geometrically the then (1961) vague concept of ill-conditioning. The diagram does *not* reveal the change in orientation of lines (hyperplanes) due to the SWOP of GE, nor does it attempt to illustrate the possible effect of instability during the BSP of GE due to poorly oriented hyperplanes rendered by the SWOP. The same can be said regarding the diagram on p. 122 of *Numerical Methods in Engineering Practices* by Al-Khafaji and Tooley.

Unlike the aforementioned published diagrams, the diagrams in this paper depict the geometries of the SWOP and the BSP of GE. However, because the geometry used in those texts to explain ill-conditioning is similar to that used in this paper to explain both phases of GE (and the subsequent error), the diagrams resemble each other and are related. All error-bound lines in Stanton's diagrams are parallel to the given lines (hyperplanes) of the original system and represent uncertainty in that system. The possible sources of this uncertainty may include initial representation error, but do not include the propagation of that error, nor the creation of other error, during the finite-precision arithmetic of GE. The error lines (planes, hyperplanes) in our diagrams are parallel to appropriate coordinate axes (planes, hyperplanes) to illustrate error propagation in the BSP, and refer not to the original system, but to the upper triangular system resulting from the SWOP of GE applied to the original system. Error in the upper triangular system includes finite-precision arithmetic error accrued during the SWOP of GE. Figures 2, 3, and 4 in this paper show how even a well-conditioned system can be affected by pivoting. The figures also illustrate the mechanism by which scaling can help to reduce errors in GE.

REMARK 4.8. While selecting pivots for a linear system, the reader might ask, "Can one exhaustively scan all possible geometric configurations and choose the best possible one in preparation for the BSP?" After presenting the concept of BSP *error multipliers* in the sequel [5] to show exactly how (and to what degree) instability may occur during the BSP, a new pivoting strategy will be introduced which "scans" for good geometric configurations. *Rook's pivoting* is a strategy which monitors and controls the error stemming from both the SWOP and the BSP of GE. Rook's pivoting is an economical (same order of magnitude) as partial pivoting, while producing a computed solution whose accuracy is comparable to that produced by complete pivoting.

REMARK 4.9. As is standard in the literature, the term "ill-conditioning" in our paper refers to the sensitivity of the computed solution to small perturbations of the original linear system. Most people use the condition number as a barometer for ill-conditioning. Some use it to obtain *a priori* bounds on the error in the computed solution, though (as Wilkinson noted) such bounds are usually coarse. On this subject, consider the philosophy of Wilkinson [4, p. 65, lines 11–25]. In the sequel, we shall show how the BSP error multipliers may provide another less expensive, more intuitive, and more sensitive measure of ill-conditioning than the condition number.

REMARK 4.10. The reader might ask, "How does this work relate to the Kaczmarz method for solving linear systems iteratively?" This very interesting question was first raised in a personal conversation with Miki Neumann (University of Connecticut, Storrs). The theme of our current work is the geometric analysis of GE, a direct method for solving linear systems. Our preliminary analysis reveals few similarities between the two methods beyond the phrase "the orientation of hyperplanes of the linear system." This work may lead to a less expensive hyperplane selection technique for the Kaczmarz method. However, the application of our geometric analysis to this iterative method is deferred.

REMARK 4.11. The geometric analysis presented in this paper should carry over well to large sparse linear systems. Such systems enjoy the property that most, if not all, hyperplanes are already parallel to almost all coordinate axes. This suggests a special pivoting strategy for sparse systems which considers the number and placement of zeros in each row.

We deeply appreciate the encouragement and advice accorded us by John Rice (Purdue University) and Jim Ortega (Virginia) during preliminary informal reviews of this work. Further, we are grateful for the special attention and careful reading of this paper by the referees, and the subsequent recommendations which improved its presentation.

REFERENCES

1 F. L. Bauer, Optimally scaled matrices, *Numer. Math.* 5:73–87 (1963).
2 G. E. Forsythe, M. A. Malcolm, and C. B. Moler, *Computer Methods for Mathematical Computations*, Prentice-Hall, 1977.
3 G. E. Forsythe and C. B. Moler, *Computer Solution of Linear Algebraic Systems*, Prentice-Hall, 1967.
4 Gene Golub and Charles Van Loan, *Matrix Computations*, 2nd ed., Johns Hopkins U.P., 1989.
5 Larry Neal and George Poole, A geometric analysis of Gaussian elimination. II, *Linear Algebra Appl.*, submitted for publication.
6 George Poole and Larry Neal, Gaussian elimination: Is scaling necessary?, *Linear Algebra Appl.*, submitted for publication.
7 John R. Rice, *Matrix Computations and Mathematical Software*, McGraw-Hill, 1981.
8 G. W. Stewart, *Introduction to Matrix Computations*, Academic, 1973.
9 Lloyd N. Trefethen, Three mysteries of Gaussian elimination, *ACM-SIGNUM Newsletter* 20, No. 4 (Oct. 1985).
10 Lloyd N. Trefethen and R. S. Schreiber, Average-case stability of Gaussian elimination, *SIAMAX* 11, No. 3, pp. 335–360, July 1990.
11 James H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon, 1965.
12 MATLAB, A software product of Math Works, Inc. and C.B. Moler.

Received 23 May 1989; final manuscript accepted 10 August 1990